Acquisition de connaissances lexicales à partir de corpus : la sous-catégorisation verbale en français

Cédric Messiant* — Kata Gábor — Thierry Poibeau*****

* *Laboratoire d'Informatique de Paris-Nord, UMR 7030 CNRS et Université Paris 13*
99, avenue Jean-Baptiste Clément — F-93430 Villetaneuse, France
cedric.messiant@lipn.univ-paris13.fr

** *Department of Language Technology, Research Institute for Linguistics*
Hungarian Academy of Sciences — Budapest, Hongrie
gkata@nytud.hu

*** *Laboratoire LaTTiCe, UMR 8094 CNRS et École Normale Supérieure*
1, rue Maurice Arnoux — F-92120 Montrouge, France
thierry.poibeau@ens.fr

RÉSUMÉ. *Nous décrivons dans cet article une méthode permettant l'acquisition d'un lexique syntaxique des verbes du français à partir de l'analyse automatique de gros corpus. Nous évaluons cette méthode par rapport à des ressources existantes et nous montrons que notre système produit automatiquement de nouvelles données qui peuvent compléter les lexiques existants. Nous montrons enfin comment la syntaxe peut aider à faire émerger des classes lexico-sémantiques, dans la lignée des travaux de Levin (1993).*

ABSTRACT. *We present in this paper a method to automatically acquire a syntactic lexicon of subcategorization frames for French verbs directly from large corpora. The method is evaluated against existing lexical resources: we show that our system is capable of producing new frames that were not previously registered. Lastly, we show that it is possible to induce lexico-semantic classes « à la Levin » (1993) from these data.*

MOTS-CLÉS : *lexique, verbe, sous-catégorisation, acquisition à partir de corpus.*

KEYWORDS: *lexicon, verb, subcategorization frames, lexical acquisition.*

1. Introduction

Les informations lexicales, c'est-à-dire les informations se rapportant aux mots et à leurs propriétés ont pris une importance considérable pour le traitement des langues ces dernières années (Danlos, 1988 ; Laporte, 2000). Il est en effet primordial pour la grammaire d'une langue de savoir comment se combinent les mots, simples ou composés, sur le plan syntaxique comme sur le plan sémantique. La finesse de l'analyse repose finalement moins sur des principes syntaxiques généraux que sur les contraintes propres à chaque élément lexical (Gross, 1975 ; Gross et Danlos, 1988). Au sein des théories lexicalistes, comme la grammaire lexicale-fonctionnelle (LFG) ou la grammaire syntagmatique guidée par les têtes (HPSG) (Abeillé, 1993), de nombreuses recherches ont été faites pour élaborer un modèle de lexique plus sophistiqué (Bresnan et Zaenen, 1990 ; Copestake, 1992 ; Koenig et Davis, 2000). De nouvelles théories visant le développement de lexiques structurés ont alors vu le jour (Levin, 1993 ; Pustejovsky, 1995).

Parallèlement, la disponibilité de corpus électroniques de grande taille a permis de mener des études sur le comportement linguistique des éléments lexicaux et les propriétés sous-jacentes qui les influencent. En effet, le comportement des mots en contexte est d'autant plus surprenant qu'il peut varier au cours du temps, d'un locuteur à l'autre, voire en fonction de la situation ou du corpus considéré. Pour prendre un exemple simple, le verbe « *essaimer* » est réputé intransitif mais il est largement employé de manière transitive dans le journal *Le Monde* (*e.g.* « *Cuba a essaimé les effets de son syncrétisme culturel au gré des modes et des engouements* », « *Il n'en a pas moins essaimé son séjour chinois de proclamations optimistes pour l'avenir [du] pays* »). Ce sont au premier chef ces types de comportements non standard (quand on les compare à un lexique général) qui nous intéressent ici[1].

Cet article porte donc sur l'acquisition automatique d'information à partir de corpus. On s'intéresse plus particulièrement à l'analyse du verbe. Prenons pour exemple une construction typique du verbe *casser* :

$$(\textit{Luc})_{\text{ARG1}}\ \textit{a cassé}\ (\textit{la vitre})_{\text{ARG2}}\ (\textit{avec un ballon})_{\text{ARG3}}.$$

De manière schématique, on peut distinguer les informations suivantes liées au verbe :

1) le nombre et la réalisation syntaxique des arguments. Par exemple, « *casser* » peut ici être considéré comme un verbe à trois arguments, le premier (dans la forme canonique du verbe) étant un groupe nominal réalisé à la gauche du verbe (le sujet), le deuxième un groupe nominal à la droite du verbe (l'objet) et le troisième un groupe prépositionnel généralement introduit par *avec* ;

1. Le Trésor de la Langue Française informatisé enregistre malgré tout quelques emplois transitifs pour *essaimer* dans la langue littéraire. Leur proportion dans *Le Monde* n'en demeure pas moins remarquable.

2) la grille thématique, qui caractérise le rôle sémantique de chacun des arguments, c'est-à-dire la relation qu'il entretient sur le plan sémantique avec le verbe. Ainsi, dans l'exemple *supra*, l'argument 1 (« *Luc* ») joue le rôle d'agent, l'argument 2 (« *la vitre* ») est le thème et l'argument 3 (« *avec un ballon* ») est l'instrument.

Notons qu'il n'y a pas de consensus sur la nature et la granularité des rôles thématiques, ni même sur ce qui constitue la grille thématique (Cruse, 1986 ; Jackendoff, 1990). Ainsi, pour le verbe *casser*, le statut de l'instrumental n'est pas fixé, on l'a considéré ici comme un argument, mais on peut aussi en faire un modifieur suivant le cadre théorique adopté.

Une alternance désigne la relation entre deux réalisations de surface d'un même prédicat, comme *Luc a cassé la vitre avec un ballon* vs *Le ballon a cassé la vitre*. Les alternances ne préservent pas toujours la grille thématique du verbe (*cf. charger le camion de foin* vs *charger le foin dans le camion*, où *camion* passe de thème à but[2]). Les alternances ont été beaucoup étudiées au niveau linguistique (Gross, 1975 ; Levin, 1993) mais elles restent très difficiles à analyser automatiquement (parmi les rares travaux s'attaquant directement à l'analyse des alternances, voir (McCarthy, 2001)). Les techniques de désambiguisation sémantique sont encore largement insuffisantes pour distinguer finement les verbes polysémiques (comme *commander*, entre *commander un soda* et *commander à Luc de faire ceci*) (Agirre et Edmonds, 2007).

L'analyse des constructions syntaxiques et l'étiquetage des rôles thématiques ont en revanche suscité de très nombreuses recherches, et nécesssitent des techniques différentes. L'étiquetage des rôles thématiques repose généralement sur une phase d'apprentissage à partir d'un corpus annoté : l'analyseur est ensuite capable d'attribuer des rôles sémantiques en fonction de configurations syntaxiques particulières (sur la question, voir les conférences CoNLL[3] (Stevenson et Carreras, 2009) et (Moreau *et al.*, 2009) pour une expérience sur le français).

L'induction de lexiques syntaxiques est un domaine de recherche plus ancien qui s'est développé à partir du début des années 1990 (Brent, 1991 ; Manning, 1993). Ce courant de recherche repose sur l'idée que les analyseurs syntaxiques non lexicalisés peuvent produire des données relativement structurées de manière massive (en général, à partir de l'analyse de corpus de plusieurs millions de mots), de telle sorte qu'il est ensuite possible de repérer des régularités de comportement et donc d'inférer des connaissances sur les constructions possibles du verbe (c'est-à-dire le nombre et la nature syntaxique des arguments, ce que l'on appelle ici schéma de sous-catégorisation, abrégé en SSC). Ces constructions sont associées « en vrac » aux lemmes verbaux : comme nous l'avons mentionné *supra*, l'état de l'art est encore insuffisant pour aller au-delà, faute notamment de techniques de désambiguisation sémantique suffisamment efficaces. L'information repérée est donc partielle et nécessite un travail manuel de tri et de validation (*cf.* section 5.1).

---

2. Sur toutes ces questions, voir le site Sémanticlopédie : `http://www.semantique-gdr.net/dico/`.
3. *Computational Natural Language Learning*.

L'hypothèse d'une corrélation entre classes syntaxiques et classes sémantiques a été posée a plusieurs reprises, notamment dans les travaux du LADL (Laboratoire d'Automatique Documentaire et Linguistique) dès les années 1970 (*cf.* (Borillo, 1971) pour les verbes symétriques, (Gross, 1975) pour les verbes psychologiques ou (Guillet et Leclère, 1992) pour les verbes locatifs). Nous nous inspirons en outre ici de B. Levin, qui a proposé de représenter le sens des verbes par le recours à des composants sémantiques (Levin, 1993 ; Levin et Rappaport Hovav, 2005). C'est en partant de cette notion que Levin arrive à fournir une description systématique des alternances. Elle présume que chaque verbe peut être décrit grâce à un noyau sémantique d'une part, et à un ensemble de composants sémantiques d'autre part. Tandis que le noyau est toujours spécifique à l'unité lexicale, les composants sémantiques sont communs à tous les verbes d'un même groupe sémantique. Les alternances syntaxiques, à leur tour, sont liées à ces composants : les verbes qui appartiennent à un même groupe sémantique (*i.e.* qui partagent les mêmes composants de sens) participent aux mêmes alternances syntaxiques.

Il existe déjà de nombreux dictionnaires à large couverture pour le français (*cf.* section 2.1) et il est évident que les méthodes automatiques n'auront ni la richesse ni la précision du résultat d'un travail minutieux mené par une équipe de linguistes[4]. En revanche, il est aisé de montrer que les mots, à l'exemple du verbe « *essaimer* » dans le journal *Le Monde*, ont des comportements variables en fonction du contexte d'usage, et donc du corpus étudié. Il paraît aujourd'hui quasi inconcevable de repérer ces comportements spécifiques par une analyse manuelle.

Partant de ce constat, les méthodes automatiques ont connu un certain succès car les techniques mises au point permettent 1) de compléter des lexiques existants en mettant le doigt sur des constructions nouvelles ou absentes des dictionnaires, 2) d'obtenir des informations sur la productivités des différentes constructions et 3) d'inférer ensuite des classes de verbes partageant les mêmes alternances, selon l'hypothèse de Levin (Levin, 1993 ; Schulte im Walde, 2000). Les informations repérées doivent ensuite être validées manuellement puis associées à des entrées lexicales pour produire des ressources vraiment utilisables. Malgré ces limites, la popularité de ce type de méthodes depuis une dizaine d'années montre leur intérêt et leur potentiel pour la mise au point de ressources à large échelle (voir l'expérience décrite dans (Kipper *et al.*, 2008)).

Cet article comporte cinq parties. Nous dressons dans un premier temps un rapide état de l'art des techniques automatiques et des ressources disponibles pour le français. Dans un deuxième temps nous décrivons la méthode d'acquisition à partir de corpus mise au point : si celle-ci repose sur l'approche développées pour d'autres langues, elle intègre aussi un certain nombre de spécificités que nous détaillons. La troisième

4. L'existence de ressources de qualité (Trésor de la Langue Française, lexique-grammaire, etc.) à large couverture d'une part, et l'absence de corpus librement disponible d'autre part, a sûrement limité en France l'intérêt pour les méthodes d'acquisition automatique par le passé. Comme nous essayons de le montrer dans la partie suivante, nous pensons que l'état des techniques est aujourd'hui tel qu'il est intéressant d'y revenir.

section présente le lexique lui-même, qui est évalué dans la section suivante : nous fournissons différentes mesures permettant d'évaluer le recouvrement par rapport à des ressources existantes et une évaluation manuelles de nouvelles constructions non répertoriées dans des ressources de référence. Dans un dernier temps, nous présentons une méthode permettant de générer des classes de verbes à partir des informations syntaxiques obtenues précédemment. Nous évaluons ces classes et discutons leur pertinence, notamment pour le repérage automatique des alternances.

2. État de l'art

Nous donnons dans un premier temps la description d'un ensemble de ressources sur le verbe pour le français. La plupart de celles-ci sont le résultat du travail manuel d'équipes de linguistes. Nous présentons ensuite les techniques d'acquisition à partir de corpus.

2.1. *Les lexiques existants pour le français*

Plusieurs ressources lexicales syntaxiques pour le français ont été développées depuis les débuts du traitement automatique des langues en France. Les objectifs de ces lexiques sont de définir, pour chaque lemme verbal donné, ses différents emplois et, pour chacun de ces emplois, son (ou ses) schéma(s) de sous-catégorisation, en spécifiant le nombre et le type des arguments, et les éventuelles informations complémentaires qui s'y rapportent.

Le lexique électronique le plus important pour le français est indéniablement le lexique-grammaire (LG) de Maurice Gross (Gross, 1975). Il rassemble les constructions syntaxiques associées à plus de 6 000 verbes dans un ensemble de tables : les lignes correspondent aux verbes, les colonnes aux différentes constructions possibles ; l'intersection d'une ligne et d'une colonne contient un signe + si la construction est possible pour le verbe, et un signe – sinon. Ce format initial a depuis été traduit sous diverses formes réputées plus facilement utilisables par les outils automatiques, comme Synlex (Gardent *et al.*, 2006) ou LGLex (Constant et Tolone, 2008).

Le dictionnaire syntaxique des verbes français (Dubois et Dubois-Charlier, 1997), mis à disposition sur le site Internet du laboratoire MoDyCo, est une classification sémantico-syntaxique des verbes manuellement construite par ces deux linguistes, dont les principes sont proches de ceux du LG. On compte dans ce dictionnaire 12 130 verbes, ce qui le rend remarquablement riche.

DicoValence (van den Eynde et Mertens, 2006) est un dictionnaire syntaxique construit manuellement dans le cadre méthodologique de l'Approche Pronominale (van den Eynde et Blanche-Benveniste, 1978). Pour identifier la valence d'un prédicat (ses dépendants et leurs caractéristiques), l'Approche Pronominale exploite la relation qui existe entre les dépendants dits lexicalisés (réalisés sous forme de syntagmes)

et les pronoms qui couvrent ces lexicalisations possibles. DicoValence comporte les schémas de sous-catégorisation de 3 738 verbes, répartis en 8 313 entrées.

Le Lefff (Sagot, 2010) est un lexique des formes fléchies du français constitué en partie par des moyens automatiques (analyse de corpus, fusion de données provenant de différentes ressources) et en partie manuellement, notamment pour la validation des entrées. Il comprend dans sa version actuelle plus de 7 000 lemmes verbaux. Le lexique est disponible sous une forme compacte (niveau intensionnel) ou sous une forme éclatée (niveau extensionnel, où chaque entrée est une forme fléchie).

Nous pouvons également mentionner d'autres ressources comme LexValf (Salkoff et Valli, 2006) dont les principes de base sont ceux des grammaires en chaîne, DiCoLAF (Mel'cuk et Polguère, 2006), centré sur la modélisation formelle des collocations et de la dérivation sémantique du français, DicoLPL (van Rullen *et al.*, 2005) ou encore le Trésor de la Langue Française informatisé (TLFI) (Dendien et Pierrel, 2003).

Les travaux de constitution de lexiques suite à un travail manuel, comme ceux présentés dans cette section, permettent d'obtenir des données relativement riches et précises. Il faut toutefois souligner la masse de travail demandé et les limites de ces lexiques : ils sont peu adaptables ou, en tout cas, leur adaptation demande un nouveau travail manuel qui est souvent incompatible avec les délais imposés par les besoins ; les maintenir et les mettre à jour demande un effort quotidien lourd et coûteux. L'arrivée conjointe de nouveaux besoins d'un côté, de corpus électroniques et d'outils de traitement relativement efficaces de l'autre, a suscité un intérêt pour des méthodes d'acquisition semi-automatiques de lexiques à partir de corpus.

2.2. *Les méthodes d'acquisition automatique de schémas de sous-catégorisation*

Des travaux sur l'acquisition d'informations de sous-catégorisation à partir de corpus brut ont été menés pour l'anglais dès le début des années 1990 (Manning, 1993 ; Brent, 1993). Ces premiers travaux étaient toutefois limités quant au nombre de verbes considérés et de SSC possibles (généralement quelques dizaines de verbes et autant de SSC). Ils reposaient par ailleurs souvent sur des heuristiques locales, sans exploiter pleinement le corpus.

Le système développé à l'Université de Cambridge (Briscoe et Carroll, 1997) est le premier a permettre une acquisition à large échelle de bonne qualité. Il a été constamment amélioré depuis, pour couvrir de nouveaux SSC ou de nouvelles parties du discours (noms, adjectifs) (Korhonen *et al.*, 2000 ; Preiss *et al.*, 2007). Il est fondé sur un analyseur de surface de l'anglais appelé RASP[5], ainsi que sur des règles d'appariement complexes entre SSC et réalisations possibles dans les textes. Il repose donc sur une énumération *a priori* des différents schémas syntaxiques visés, ce qui facilite la tâche mais ne permet pas la découverte de structures complètement nouvelles.

5. http://www.informatics.sussex.ac.uk/research/groups/nlp/rasp/

C'est pourquoi nous avons choisi de ne pas spécifier une telle liste *a priori* dans notre approche, ce qui la rend plus portable.

Pour le français, P. Chesley et S. Salmon-Alt ont mené une étude exploratoire sur 104 verbes fréquents qui leur ont permis de repérer 27 SSC différents (Chesley et Salmon-Alt, 2006). Par la suite, dans le cadre du projet ANR Passage (`http://atoll.inria.fr/passage/`), C. Gardent a mené une expérience portant sur un nombre beaucoup plus important de verbes en partant d'un corpus de 100 millions de mots[6]. Ce corpus a été ensuite analysé au moyen de l'analyseur syntaxique TagParser mis au point par G. Francopoulo (Francopoulo, 2005). Le repérage de régularités au niveau des compléments du verbe permet d'inférer des SSC pour chaque verbe, suivant une stratégie proche de celle de l'équipe de Cambridge. Le lexique résultant, EasyLex, est disponible sur le portail TALC (`http://talc.loria.fr/Easylex.html`).

Tous les systèmes mentionnés ici obtiennent des performances qui peuvent apparaître relativement médiocres. Le rappel dépasse rarement 0,65 et la précision est en général un peu meilleure. Qu'est-ce que cela signifie ? Le fait que le système ne permet pas d'acquérir un SSC donné à partir d'un corpus précis ne signifie pas obligatoirement qu'il y a erreur : il peut tout simplement s'agir d'un emploi du verbe absent du corpus. L'intérêt et l'« utilisabilité » de la méthode doivent donc être mesurés : quelles sont les performances réelles ? Quels sont les cas d'usage possibles ? Nous essayons d'aborder cette question de front, alors qu'elle a paradoxalement été peu traitée jusqu'ici par les auteurs sus-cités[7].

Signalons enfin une autre façon d'appréhender l'extraction de SSC, en partant directement d'un corpus arboré. De nombreuses expériences ont été faites dans ce cadre pour l'anglais (O'Donovan *et al.*, 2005), et une expérience similaire a été faite sur le français : TreeLex (Kupsc, 2007) est un lexique de sous-catégorisation verbale pour le français contemporain extrait automatiquement du corpus arboré de Paris 7 (Abeillé *et al.*, 2003). Il contient à peu près 2 000 lemmes verbaux et 180 SSC (moyenne de 2,09 schémas par lemme). Même si l'on peut aussi parler d'acquisition dans ce cas, il s'agit en fait d'une approche très différente de la nôtre. Les corpus arborés sont excessivement rares : la méthode est donc peu portable et ne permet pas de traiter du corpus « tout-venant ». TreeLex est toutefois intéressant pour nous car il s'agit d'un lexique en principe correct (car dérivé d'un corpus annoté et validé manuellement), et acquis à partir du journal *Le Monde*. Nous nous servirons donc de TreeLex comme point de comparaison pour évaluer nos résultats.

6. Il s'agit du Corpus Passage Court – CPC `http://atoll.inria.fr/passage/ressources.en.html`.

7. A. Korhonen, dans sa thèse (Korhonen, 2002), fait une analyse très complète de ses résultats par rapport à un *gold standard*. Il faut toutefois noter la difficulté de cerner la notion de *gold standard* pour des lexiques : comment définir la complétude d'un lexique ? Celle-ci doit-elle être mesurée dans l'absolu, par rapport à un corpus, à une tâche ? (Poibeau et Messiant, 2008)

3. Acquisition automatique de schémas de sous-catégorisation : le système ASSCi

ASSCi est le système d'acquisition automatique de schémas de sous-catégorisation que nous avons développé pour l'analyse des verbes français (Messiant, 2008). Après une présentation de l'architecture globale d'ASSCi, nous présentons les outils utilisés pour les prétraitements puis les trois modules qui composent le système proprement dit : l'extracteur de pré-schémas de sous-catégorisation locaux, le constructeur de schémas candidats et le filtre de schémas non pertinents.

3.1. *Architecture générale d'ASSCi*

L'architecture d'ASSCi est inspirée des principaux travaux récents en matière d'acquisition automatique de SSC à partir de corpus (Preiss *et al.*, 2007). Les quatre étapes principales de ce modèle sont :

1) une phase de prétraitement durant laquelle les phrases sont annotées à travers une analyse de surface. Dans ASSCi, le corpus brut est lemmatisé et annoté par TreeTagger puis analysé par l'analyseur de surface SYNTEX ;

2) l'identification des verbes et de leurs compléments parmi ces données annotées. Dans ASSCi, ce rôle est tenu par l'extracteur de pré-schémas de sous-catégorisation locaux qui extrait pour chaque phrase les informations utiles pour constituer les futurs SSC ;

3) les schémas de sous-catégorisation candidats sont ensuite inférés à partir de ces informations. Dans ASSCi, le constructeur de schémas candidats s'occupe de rassembler les SSC observés en corpus pour chaque verbe ;

4) un filtrage vise à distinguer les SSC erronés des SSC corrects pour chaque verbe. Dans ASSCi, le filtre des SSC non pertinents est fondé sur des méthodes statistiques.

À l'issue du processus d'acquisition, le système produit un lexique composé de couples verbes–SSC ainsi que d'informations statistiques et lexicales associées à ces couples (le lexique est présenté dans la section 4).

3.2. *Prétraitements*

Les étapes préalables à la tâche d'acquisition sont la lemmatisation, l'analyse morphosyntaxique et l'analyse syntaxique de surface. Au regard des outils disponibles pour le français et des performances de ceux-ci, nous avons décidé d'utiliser l'analyseur syntaxique SYNTEX. Cet analyseur repose sur les annotations de l'analyseur morphosyntaxique TreeTagger. Cette section présente ces deux outils ainsi qu'un exemple d'annotation et d'analyse.

3.2.1. *Annotation morphosyntaxique : TreeTagger*

TreeTagger est un outil de lemmatisation et d'annotation en parties du discours (Schmid, 1994)[8]. TreeTagger fournit en outre des outils pour la segmentation en phrases et en mots[9], puis associe une étiquette morphosyntaxique à chaque élément de la phrase[10]. Enfin, TreeTagger s'occupe de la lemmatisation des mots de la phrase.

L'un des avantages de TreeTagger, outre sa robustesse et son efficacité, est son ouverture : il est possible d'ajouter des traitements en amont (par exemple, faire à sa place le découpage en unités de traitement (*tokenisation*) ou l'étiquetage). Ainsi, des règles et des lexiques de reconnaissance des unités syntaxiques complexes (*e.g.* locutions prépositionnelles) ont été ajoutés pour SYNTEX (Bourigault *et al.*, 2005). Il est également possible d'intégrer dans la chaîne de traitement des règles de *tokenisation* et de pré-étiquetage spécifiques au corpus à analyser, ce qui est fondamental lorsque l'étiqueteur doit traiter des données non standard (codes de produits, nomenclature d'éléments chimiques, etc.). L'analyseur syntaxique a enfin la possibilité de faire des retours en arrière sur l'étiquetage et de modifier les étiquettes attribuées par TreeTagger.

3.2.2. *Analyse syntaxique : SYNTEX*

Le corpus est ensuite analysé par SYNTEX, analyseur syntaxique en dépendances développé par Didier Bourigault (Bourigault *et al.*, 2005 ; Bourigault, 2007). SYNTEX réalise une analyse syntaxique en dépendances : les principales relations syntaxiques reconnues par l'analyseur sont les suivantes : sujet, complément d'objet direct, complément prépositionnel (de nom, de verbe et d'adjectif), antécédence relative (*i.e.* antécédent des pronoms relatifs), modification adjectivale (épithète, attribut) et subordination. Chaque élément de la phrase est annoté par ses relations de recteur ou de « régi » avec les autres éléments (par exemple, dans le cas d'un verbe transitif, le verbe est « recteur » du sujet et de l'objet ; inversement, ces derniers sont « régis » par le verbe).

Pour annoter les éléments de la phrase, SYNTEX applique différents modules de reconnaissance de relations syntaxiques en série : chaque module prend en charge une relation syntaxique particulière et l'entrée de chaque module est la sortie du module qui le précède. Cependant, des retours en arrière sont possibles dans la chaîne de traitement et un module peut remplacer l'étiquette apposée par un module qui l'a précédé, si nécessaire. Ce fonctionnement rend malgré tout essentiel le choix de l'ordre d'exécution des modules.

SYNTEX repose fondamentalement sur des règles et des procédures d'apprentissage endogène, même si des informations lexicales ponctuelles sont utilisées en

8. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`
9. Le fichier de paramétrage du segmenteur pour le français a été fourni par Michel Généreux.
10. Les fichiers de paramètrage de TreeTagger en français sont fournis par Achim Stein : `http://www.uni-stuttgart.de/lingrom/stein/forschung/resource.html`.

fonction des besoins. L'analyseur reste toutefois peu lexicalisé : il dispose simplement de quelques listes, par exemple pour les locutions prépositionnelles. SYNTEX n'utilise pas de lexique de sous-catégorisation à proprement parler ; les rattachements se font sur la base de probabilités acquises à partir d'un grand corpus, notamment en ce qui concerne les groupes prépositionnels (et, du coup, arguments et modifieurs sont rattachés au verbe sans qu'il soit fait de distinction entre eux). Une meilleure prise en compte de la sous-catégorisation permettrait probablement d'améliorer les performances de l'analyseur en considérant l'intégralité de la structure argumentale et plus seulement des relations locales.

Le choix de SYNTEX comme analyseur syntaxique pour notre système d'acquisition automatique de SSC s'explique à la fois par ses propriétés, ses performances sur les corpus écrits et par sa disponibilité. Les résultats de SYNTEX lors de la campagne d'évaluation EASY en 2007[11] montrent que l'analyseur obtint alors les meilleurs résultats en précision et en F-mesure sur les corpus écrits[12]. De plus, la robustesse de SYNTEX garantit son adaptabilité, du moins sur les textes qui respectent les normes syntaxiques de l'écrit standard. La réutilisation de notre système d'acquisition sur des corpus de domaines particuliers (médecine, droit...) ne devrait donc pas poser de problème.

Néanmoins, il existe encore une part non négligeable d'erreurs dues à des faiblesses de l'étiquetage morphosyntaxique ou de l'annotation des relations. La plupart de ces erreurs se répercuteront dans le processus d'acquisition des SSC et devront être traitées ou filtrées au cours du processus.

3.2.3. *Exemple*

L'exemple suivant présente l'annotation produite par TreeTagger et l'analyse faite par SYNTEX pour la phrase[13] :

Il les lui reproche au nom du Sartre qu'il aime.

Dans le tableau 1, la colonne de gauche correspond à la sortie de TreeTagger tandis que la colonne de droite correspond à la sortie de SYNTEX. TreeTagger fournit des étiquettes morphosyntaxiques et les lemmes correspondant aux éléments de la phrase. SYNTEX associe les informations suivantes (séparées par des "|") à chaque élément de la phrase : la catégorie morphosyntaxique de l'élément (`VCONJS`) ; le lemme (`reprocher`) ; l'élément tel qu'il se trouve dans la phrase d'origine (`reproche`) ; l'identifiant de l'élément dans la phrase (`4`) ; la liste des recteurs de l'élément ; la liste des dépendants (*i.e.* éléments régis) de l'élément (`SUJ;1,OBJ;2,PREP;3,PREP;5`).

---

11. `http://www.limsi.fr/Recherche/CORVAL/easy/`

12. L'ensemble des résultats de SYNTEX à la campagne EASY sont disponibles sur la page consacrée au logiciel : `http://w3.erss.univ-tlse2.fr/membres/bourigault/syntex.html`. On remarquera le différentiel de performance important entre SYNTEX et ses poursuivants en terme de précision sur les corpus écrits lors de cette campagne d'évaluation.

13. Nous choisissons volontairement une phrase non canonique, à l'image de la majorité des phrases à traiter dans notre corpus.

```
PRO:PER    il          | Pro|il|Il|1|SUJ;4|
PRO:PER    la|le       | Pro|le|les|2|OBJ;4|
PRO:PER    lui         | Pro|lui|lui|3|PREP;4|
VER:subp   reprocher   | VCONJS|reprocher|reproche|4||SUJ;1,OBJ;2,PREP;3,PREP;5
PRP:det    au          | Prep|au nom de|au nom du|5|PREP;4|NOMPREP;6
NOM        nom         |
PRP:det    du          |
NOM        Sartre      | NomPrXXInc|Sartre|Sartre|6|NOMPREP;5|
PRO:REL    que         | CSub|que|qu'|7||COMP;9
PRO:PER    il          | Pro|il|il|8|SUJ;9|
VER:pres   aimer       | VCONJS|aimer|aime|9|COMP;7|SUJ;8
SENT       .           | Typo|.|.|10||
```

Tableau 1. *Analyse d'une phrase avec TreeTagger et Syntex*

Fonction		Catégorie	
SUJ	sujet	SN SINF	syntagme nominal syntagme infinitif
REF	forme pronominale	refl	pronom
OBJ	objet	SN SINF PropSub	syntagme nominal syntagme infinitif proposition subordonnée
ATTS	attribut du sujet	SA SN SINF	syntagme adjectival syntagme nominal syntagme infinitif
ATTO	attribut de l'objet	SA	syntagme adjectival
A-OBJ	objet indirect gouverné par *à*	SP<à+SN> SP<à+SINF>	syntagme prépositionnel syntagme prépositionnel
DE-OBJ	objet indirect gouverné par *de*	SP<de+SN> SP<de+SINF>	syntagme prépositionnel syntagme prépositionnel
P-OBJ	objet indirect gouverné par la préposition *prep*	SP<*prep*+SN> SP<*prep*+SINF>	syntagme prépositionnel syntagme prépositionnel

Tableau 2. *Cas traités par le constructeur de schémas candidats*

3.3. *Extracteur de pré-schémas de sous-catégorisation locaux*

Le premier module extrait des « pré-schémas de sous-catégorisation » (pré-SSC) : à chaque occurrence de verbe conjugué du corpus est associé un pré-schéma constitué du verbe à l'infinitif et de la liste de ses compléments. Pour chaque complément, le module fournit le lemme de sa tête, sa catégorie morphosyntaxique et sa fonction syntaxique. L'extracteur associe l'un des couples (fonction et catégorie) d'étiquettes présentées dans le tableau 2 à chacun des compléments du verbe. Les phrases reconnues par SYNTEX comme étant à la forme passive sont marquées par l'ajout du code `PASSIF`.

Par exemple, pour le verbe issu de la partie de phrase « *il les lui reproche au nom du Sartre qu'il aime* », l'extracteur produit le pré-SSC suivant :

```
0100.anasynt!d686339p6_2!21
REPROCHER+reprocher
[P-OBJ:SP<au_nom_de+SN>:Sartre, SUJ:SN:il, OBJ:SN:le, A-OBJ:SP<à+SN>:lui]
```

Le pré-schéma est composé de trois éléments : l'identifiant du pré-schéma (constitué du nom du fichier dont est issu le pré-SSC, de l'identifiant de la phrase dans le fichier et de l'identifiant du verbe concerné dans la phrase), du verbe concerné par le pré-schéma et de ses compléments (ici, *reprocher* a quatre compléments, de type `P-OBJ`, `SUJ`, `OBJ`, `A-OBJ`).

3.4. *Constructeur de schémas de sous-catégorisation candidats*

Le constructeur de schémas utilise les informations contenues dans les pré-SSC locaux pour « construire » les SSC candidats qui constituent un lexique de sous-catégorisation non filtré. Contrairement à la plupart des méthodes d'acquisition de SSC employées jusqu'alors, comme dans (Preiss *et al.*, 2007), ASSCi ne dispose pas d'une liste de schémas prédéfinie. Ce choix s'inscrit dans notre volonté d'acquérir des schémas sans *a priori*, pour faire émerger du corpus les SSC correspondant à l'usage. Il n'y a d'ailleurs pas de raison qu'une liste de schémas existe si l'on ne dispose pas déjà d'une ressource et il est clair que la constitution d'une liste de schémas possibles pour une langue donnée n'est pas une tâche triviale. Ce choix permet même d'observer pour certains verbes des constructions qu'il était difficile de prévoir *a priori*, surtout quand on travaille sur corpus spécialisé.

Les SSC candidats ne concernent pas une seule occurrence en corpus mais un ensemble d'occurrences. Le constructeur a aussi pour rôle de comptabiliser le nombre d'occurrences de chaque schéma de sous-catégorisation pour chaque verbe ainsi que de calculer leur fréquence relative, c'est-à-dire la fréquence du SSC relativement au verbe. Ces informations de fréquence seront par la suite utilisées par le module de filtrage (voir section 3.5).

Les éléments des SSC sont d'abord ordonnés selon leur fonction, dans l'ordre suivant : sujet, objet, groupe prépositionnel gouverné par *à*, groupe prépositionnel gouverné par *de*, autres groupes prépositionnels, attribut du sujet, attribut de l'objet. Cette normalisation permet de s'affranchir des réalisations de surface. En effet, les phrases contenant des pronoms, des inversions ou des extrapositions n'ont pas une structure canonique et l'on souhaite s'abstraire de ces variations de surface.

Les compléments gouvernés par une préposition ne pouvant gouverner un argument sont ensuite supprimés. La liste des prépositions concernées est issue du lexique PREPLEX, un lexique de prépositions du français construit en fusionnant les informa-

tions contenues dans différents lexiques disponibles[14] (Fort et Guillaume, 2007). Le module supprime également les compléments « doublons » des pré-SSC. On considère que deux compléments sont doublons lorsqu'ils sont strictement similaires par leur fonction et leur catégorie. En effet, la plupart du temps, ces doublons sont dus à des erreurs d'analyse syntaxique et les constructions contenant deux compléments similaires sont excessivement rares en français.

Après ces traitements chargés de la normalisation et de la constitution des SSC, le constructeur de SSC calcule le nombre d'occurrences de chaque couple verbe–schéma et sa fréquence relative pour le verbe considéré, comme suit :

$$freq_rel(verbe_i, ssc_j) = \frac{|verbe_i, ssc_j|}{|verbe_i|}$$

Ces données seront ensuite utilisées lors de l'étape de filtrage.

3.5. *Filtre des schémas non pertinents*

L'acquisition se termine par une phase de filtrage : en effet, les schémas candidats extraits par le constructeur sont bruités car ils sont parfois construits à partir d'étiquetages ou d'analyses syntaxiques incorrects. Il se peut également qu'une règle de construction produise des SSC incorrects à cause d'une généralisation ou d'une normalisation trop importante. On constate enfin la présence de modifieurs dans certains schémas, qu'il faut donc filtrer.

Le filtre examine les SSC candidats issus du constructeur et compare leur fréquence relative à un seuil déterminé empiriquement à l'aide d'un jeu de test. Si la fréquence du schéma est inférieure au seuil, il est rejeté par le module. Cette méthode est inspirée de la stratégie de filtrage utilisée par (Korhonen, 2002), fondée sur le maximum de vraisemblance (*maximum likelihood estimates*). Korhonen compare le maximum de vraisemblance avec deux autres techniques de filtrage : la log-vraisemblance (*log likelihood ratio*, (Korhonen *et al.*, 2000)) et la loi binomiale, souvent utilisée (*binomial hypothesis testing*, *cf.* (Brent, 1991 ; Briscoe et Carroll, 1997)) : c'est finalement le filtrage par un jeu de seuils estimé en fonction des données qui est la méthode donnant les meilleurs résultats.

La sortie de ce module est une liste de SSC qui constitue le lexique de sous-catégorisation. Par exemple, pour la phrase « *il les lui reproche au nom du Sartre qu'il aime* », le filtre produit le SSC suivant : [SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>]. La fréquence relative du schéma candidat (voir section 3.4) est égale à 0,256, c'est-à-dire supérieure au seuil fixé empiriquement à 0,1. L'entrée est donc retenue dans le lexique.

Afin d'améliorer cette méthode, nous avons déterminé des seuils différenciés pour les SSC ne contenant qu'un sujet (schéma INTRANS, c'est-à-dire [SUJ:SN]) et les

14. http://loriatal.loria.fr/Resources.html

schémas avec un marqueur de forme pronominale. En effet, lorsque le rattachement d'un élément à un autre est incertain, SYNTEX préfère ne pas les lier. Certaines phrases complexes (avec des incises, notamment) produisent des analyses erronées et le schéma résultant est souvent le schéma INTRANS. C'est pourquoi le système utilise dans ce cas un seuil plus élevé. La même stratégie est également utilisée en présence de compléments pronominaux, souvent difficiles à distinguer lors de l'analyse, ce qui entraîne des erreurs assez fréquentes.

Une difficulté persistante est l'amélioration de la distinction argument/modifieur. Comme dans la plupart des travaux existants (Briscoe et Carroll, 1997), nous comptons sur le fait que les SSC comprenant des modifieurs sont plus variés en corpus que les SSC comprenant uniquement des arguments (les modifieurs peuvent être introduits par un grand nombre de prépositions car ils ne sont pas directement contraints par le verbe). L'un des rôles de l'étape de filtrage est donc de repérer les schémas incluant des modifieurs puis de les analyser pour ne garder que les arguments.

Nous avons mis en place un processus de réduction de ces schémas avec modifieur : lorsqu'un schéma contenant au moins un complément prépositionnel est rejeté par le filtrage parce qu'il est trop rare, on essaie de le ramener à un schéma moins complexe, c'est-à-dire sous-catégorisant un complément prépositionnel de moins. On espère ainsi éliminer un modifieur et ramener le schéma à un SCC valide pour le verbe (sur la base de sa fréquence relative). Les schémas sont donc logiquement traités par ordre décroissant de longueur (*i.e.* leur nombre de compléments) et les fréquences relatives recalculées pour tenir compte de l'étape de filtrage.

Soit par exemple la phrase *Jean boit un café à la terrasse*. Cette phrase permet d'identifier un SSC où le groupe prépositionnel introduit par *à* est un complément potentiel, mais comme la fréquence de ce SSC est inférieure au seuil, le schéma est « réduit », ce qui permet d'identifier ici un emploi transitif `[SUJ:SN, OBJ:SN]` du verbe *boire*.

4. Expérience : acquisition de LEXSCHEM

L'application d'ASSCi sur un gros corpus journalistique (LM10) a permis d'acquérir un lexique de sous-catégorisation pour le français : LEXSCHEM[15].

4.1. *Le corpus LM10*

Le corpus choisi pour l'acquisition de LEXSCHEM est un corpus composé des articles du quotidien *Le Monde* sur 10 ans (1991-2000, 200 millions de mots) obtenu

15. Il est question ici de la dernière version de LEXSCHEM (version 3 disponible sur le Web, `http://www-lipn.univ-paris13.fr/~messiant/lexschem.htm`), qui est la plus exacte à l'heure à laquelle nous rédigeons cet article (février 2010).

auprès de l'agence *ELRA*. Ce choix comporte un double avantage : il s'agit d'un corpus de référence suffisamment « propre » pour limiter les erreurs d'analyse (*SYNTEX* a obtenu sur ce type de corpus une précision de 0,76 et un rappel de 0,58 lors de la campagne *EASY*) (Bourigault, 2007). Le corpus est par ailleurs de type journalistique, ce qui assure à la fois une certaine homogénéité quant au style, et une certaine hétérogénéité quant aux thèmes abordés (*Le Monde* traite aussi bien de sport que de politique, d'économie que de vie quotidienne).

4.2. LEXSCHEM

LEXSCHEM comprend 10 928 entrées, correspondant à des combinaisons verbes–SSC différentes. Ces entrées concernent 5 261 verbes et 112 SSC distincts. Chaque entrée du lexique est composée du verbe concerné, du SSC, du nombre d'occurrences de l'entrée en corpus ainsi que des têtes argumentales et des identifiants des phrases d'où l'entrée a été extraite.

Par exemple, la phrase « *il les lui reproche au nom du Sartre qu'il aime.* » a permis (avec d'autres phrases du corpus), d'inférer le schéma candidat :

```
<ID>            2610
<VERB>          REPROCHER+reprocher
<VERB_NB_OCC>   9757
<SCF>           [SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>]
<NB_OCC>        2128
<VERB_NB_SCF>   118
<REL_FREQ>      0.218099825766117
<SEQ_ID>        0100.anasynt!d6863p6_2!4, 0100.anasynt!d6835p2_7!9, ...
<NB_ARGS>       3
<ARG0>          il,on,...
<ARG1>          le,manque,...
<ARG2>          lui,secrétaire,...
<PASS>          oui
```

Les champs disponibles pour chaque entrée sont :

– ID : l'identifiant de l'entrée ;

– VERB : le verbe concerné par l'entrée ;

– SCF : le schéma de sous-catégorisation concerné par l'entrée ;

– NB_OCC : le nombre d'occurrences de l'entrée dans le corpus ;

– VERB_NB_OCC : le nombre d'occurrences du verbe dans le corpus ;

– VERB_NB_SCF : le nombre de SSC pour ce verbe ;

– REL_FREQ : la fréquence relative de l'entrée ;

– SEQ_ID : les identifiants des séquences de l'analyse SYNTEX qui ont produit cette entrée ;

– NB_ARGS : le nombre d'arguments du SSC ;

– ARG*n* : les lemmes têtes du n-ième argument ;

– PASS : la possibilité de transformation à la voix passive.

Dans le SSC, les arguments sont présentées sous forme de liste entre crochets, séparés par des virgules. Par exemple, le schéma `[SUJ:SN, OBJ:SN, A-OBJ:SP<à+SN>]` contient trois arguments : le sujet (`SUJ:SN`), le complément d'objet (`OBJ:SN`) et le complément prépositionnel régi par *à* (`A-OBJ:SP<à+SN>`). Remarquons que le complément prépositionnel régi par *au nom de* (`P-OBJ:SP<au_nom_de+SN>`) n'apparaît pas dans le SSC retenu (car cette préposition ne peut introduire un argument, suivant les données issues de PREPLEX).

LEXSCHEM est disponible et consultable *via* une interface graphique à l'url suivante : `http://www-lipn.univ-paris13.fr/~messiant/lexschem.htm`. La version la plus exacte est actuellement la version 3 mais des mises à jour sont effectuées régulièrement en fonction des améliorations apportées aux outils. Signalons enfin différentes sous-versions du lexique, en fonction de la stratégie de filtrage et de l'information retenue (les seuils de filtrage peuvent être plus ou moins élevés, le lexique peut garder en mémoire les têtes nominales des arguments, etc.). Ces sous-versions sont utiles pour répondre à des besoins variés : les applications de TAL reposeront en majorité sur la version filtrée la plus exacte mais le calcul de classes de comportement lexico-sémantiques peut profiter de la version non filtrée et éventuellement des connaissances sur le contenu lexical des arguments du verbe.

Le lexique est disponible dans un format standard facilement traduisible sous différentes formes, notamment vers le format EASY (section 5.1) ou LMF (Francopoulo *et al.*, 2006).

5. Évaluation

Cette partie porte sur l'évaluation de LEXSCHEM, le lexique obtenu à partir du corpus LM10. Nous le comparons d'abord avec d'autres lexiques du français, avant d'examiner plus en détail la nature des informations qu'il contient et l'importance du type de corpus utilisé lors de l'acquisition. Nous proposons enfin une typologie des erreurs du système.

5.1. *Évaluation quantitative : comparaison avec d'autres ressources*

La manière la plus classique d'évaluer une ressource acquise automatiquement est de la comparer à d'autres ressources. Il s'agit de l'approche généralement adoptée (Preiss *et al.*, 2007).

5.1.1. *Ressources utilisées et format commun*

À des fins de comparaison et d'évaluation, C. Gardent a unifié un certain nombre de ressources dans un format pivot dans le cadre du projet TALC (traitement automatique des langues et des connaissances)[16]. Le processus n'est évidemment pas sans poser problème : les ressources concernées ne sont pas fondées sur la même théorie, elles n'ont pas le même objectif et les informations n'y sont pas toujours codées de façon explicite. De plus, l'alignement des données entraîne forcément un appauvrissement des ressources.

Ce mode d'évaluation reste malgré tout intéressant, à notre avis, au moins à des fins de comparaison. Nous sommes donc partis de deux lexiques de sous-catégorisation pour le français : TreeLex et DicoValence, choisis pour leurs spécificités. En effet, comme nous l'avons vu dans la section 2, ces deux ressources ne sont pas comparables, même si des similarités existent.

TreeLex est issue d'un corpus annoté (le corpus arboré de Paris 7), DicoValence est le fruit d'un travail manuel. Les deux lexiques reposent donc, directement ou indirectement, sur un important travail de description linguistique préalable. Il faut par ailleurs noter que DicoValence est structuré autour de la notion d'entrée lexicale, tandis que TreeLex fournit une liste de schémas de sous-catgéorisation « à plat » (c'est-à-dire que les constructions sont directement associées à un lemme verbal et non à une unité lexicale reflétant les distinctions de sens, comme dans un dictionnaire usuel). De ce point de vue, LEXSCHEM est très proche de TreeLex ; l'absence d'entrées lexicales en tant que telles est certes une limite, mais ceci n'est pas rédhibitoire s'il s'agit de fournir des ressources à un analyseur syntaxique ou s'il s'agit de fournir des données au linguiste qui se charge ensuite de les valider.

5.1.2. *Conversion de* LEXSCHEM

Pour effectuer la comparaison des lexiques, nous reprenons les données au format EASY produites par le LORIA et nous avons évidemment converti LEXSCHEM dans ce format. Les compléments prépositionnels y sont réduits en trois catégories : `A-OBJ` (compléments en « *à* »), `DE-OBJ` (compléments en « *de* ») et `P-OBJ` (les compléments gouvernés par toutes les autres prépositions). Ce processus de normalisation est nécessaire (le détail des prépositions est absent dans TreeLex et n'est pas systématique dans DicoValence) mais provoque un appauvrissement certain des informations contenues dans le lexique. Toutefois, la phase de validation manuelle (section suivante) permet ensuite de vérifier que les schémas comprenant des `P-OBJ` concernent les bonnes prépositions.

16. Les lexiques au format EASY peuvent être trouvés à l'adresse suivante : `http://talc.loria.fr/Dicovalence-Easy.html`. Tous les lexiques mentionnés dans cette section ont été consultés en ligne le 20 février 2010.

	TreeLex	DicoValence
Nombre total de SSC dans la ressource	3 570	4 282
Nombre de SSC communs avec LEXSCHEM	2 181	2 563
Recouvrement	61,1 %	59,9 %
Nombre de SSC nouveaux dans LEXSCHEM	1 440	1 058

Tableau 3. *Nombre de SSC communs et nouveaux dans* LEXSCHEM *par rapport aux deux ressources de référence (pour 1 583 verbes)*

5.1.3. *Résultats et discussion*

Les SSC des 1 583 verbes communs aux trois ressources ont été comparés. Le tableau 3 montre les résultats obtenus[17]

On ne remarque pas de différences significatives au niveau du recouvrement entre les deux ressources de référence, TreeLex et DicoValence. La seule différence remarquable concerne les nouveaux SSC, plus nombreux quand on compare LEXSCHEM avec TreeLex qu'avec DicoValence. Ceci s'explique probablement par le plus grand nombre de SSC dans DicoValence (donc la meilleure couverture de cette ressource ; TreeLex a été acquis sur un corpus de taille moyenne – 1 million de mots – qui ne contient qu'un sous ensemble des SSC caractéristiques du français).

Une part non négligeable des SSC des ressources de référence n'est pas retrouvée par notre méthode. Ceci est dû à la stratégie de filtrage qui élimine les SSC les moins fréquents. Il s'agit d'un biais assez courant des techniques statistiques, qu'il est difficile de corriger au niveau du filtrage lui-même : des techniques de filtrage plus sophistiquées n'ont pas montré d'amélioration significative (Korhonen, 2002). On s'aperçoit en revanche lors de l'analyse manuelle qu'un nombre non négligeable de SSC manquants pourrait être inféré, ou au moins proposé au linguiste à partir des SSC effectivement repérés : une bonne partie des SSC manquants sont en fait des formes réduites de SSC complexes (du fait du caractère optionnel dans les réalisations de surface de la plupart des arguments). Ainsi, le système a inféré que le verbe *donner* pouvait être employé avec deux compléments (*Les anticolonialistes de ce bord-là donnaient une dimension morale à leur choix*) ou un complément d'objet direct seul (*Il donne l'alerte*). Le système peut assez sûrement proposer dans ce type de cas la construction avec un complément d'objet indirect seul (*Il faut donner au denier du culte*), surtout si quelques exemples ont été trouvés en corpus. Il est donc possible d'améliorer la couverture en introduisant un processus de ce type en post-traitement, ce qui permet de réduire de plus de 50 % le nombre de SSC manquants. On arrive

17. Nous ne mentionnons pas ici les scores obtenus avec Synlex au format EASY car ceux-ci ne nous semblent pas représentatifs des résultats qui pourraient être obtenus à partir du LG complet, Synlex ayant été élaboré à partir du sous-ensemble du LG publiquement disponible (même si nos résultats sont cohérents avec ceux obtenus par C. Gardent sur les mêmes données).

alors à une couverture plus satisfaisante de LEXSCHEM par rapport aux ressources existantes.

Le tableau 3 montre l'apport possible de LEXSCHEM par rapport aux deux ressources de référence (sans prise en compte du post-traitement que nous venons de présenter). Les résultats doivent être complétés par une analyse manuelle des SSC obtenus afin de vérifier la pertinence des nouveaux schémas trouvés et la qualité des ressources de référence par rapport à notre tâche. Une évaluation manuelle sur 150 verbes a alors été menée par deux annotateurs indépendants. Cette évaluation a révélé que 108 nouveaux SSC valides pouvaient être ajoutés à TreeLex et 75 à DicoValence[18] (plus de la moitié des entrées nouvelles proposées ont été validées par les annotateurs). Ceci montre selon nous la capacité de notre système (et plus généralement des méthodes automatiques) à assister le linguiste lors de l'élaboration d'un lexique.

5.1.4. *Le cas des constructions pronominales*

Les constructions pronominales sont intéressantes parce qu'elles posent des problèmes de codage complexes, qui sont résolus de manières différentes suivant les lexiques considérées. Il s'agit en outre d'un phénomène massif en français, qui concerne la plupart des lemmes verbaux.

Prenons le cas du verbe *confondre*. Pour ce verbe, LEXSCHEM propose les SSC suivants (l'étiquette *Refl* est utilisée pour toutes les constructions pronominales, indépendamment de la valeur sémantique du pronom) :

SSC	Exemple
[SUJ, OBJ]	*Ces gentils faux bénévoles confondaient la chose et son slogan.*
[SUJ, Refl]	*Si les oiseaux avaient la télé, il y a longtemps que leurs chants se confondraient.*
[SUJ, OBJ, SP<avec>]	*Un chasseur a été tué par erreur, par un compagnon qui l' a confondu avec un sanglier.*
[SUJ, Refl, SP<avec>]	*Le roi est le patron du makhzen, mais il ne se confond pas avec lui.*

Les constructions pronominales sont souvent couvertes de manière partielle dans les différents lexiques syntaxiques du français (voir section 2.1), y compris dans des lexiques par ailleurs quasi exhaustifs. Pour le verbe *confondre*, DicoValence est le plus précis dans la mesure où le lexique encode non seulement les formes pronominales mais précise en outre la valeur sémantique des arguments (entrées 18280, 18290, 18300, 18305 et 18310). Le lexique-grammaire (et les versions dé-

18. Il faut toutefois noter que c'est à dessein que DicoValence et TreeLex ne sont pas exhaustifs : les concepteurs de DicoValence ont par exemple volontairement limité leur lexique aux 3 500 verbes les plus fréquents du français. Il n'empêche que les applications de traitement des langues exigent des dictionnaires aussi complets et précis que possible et, dans ce cadre, les méthodes (semi-)automatiques peuvent se révéler précieuses.

rivées comme LGLex) ne mentionne pas les constructions pronominales du verbe *confondre* mais précise les réalisations possibles des arguments (entrée 4_114 et 32H_153 de LGLex)[19]. TreeLex ne mentionne qu'une seule construction, non pronominale (`[SUJ:NP,OBJ:NP]`).

Dans ce cas précis, les données de LEXSCHEM sont intéressantes et pourraient compléter certains des lexiques examinés ci-dessus. Il va de soi que pour d'autres cas moins favorables, ce serait les autres ressources qui seraient les mieux armées. Un travail reste donc nécessaire pour combiner les ressources et les diverses informations disponibles. Ceci est particulièrement vrai des formes pronominales : il est rapidement nécessaire de quitter le cadre purement syntaxique pour déterminer leur valeur sémantique (réfléchi, réciproque, passif, « pseudo-se », etc.). Seul un travail manuel peut fournir des informations fines à ce niveau ; une stratégie possible est la fusion de sources de connaissances complémentaires : c'est par exemple la stratégie présentée pour compléter le Lefff sur cet aspect (Sagot et Danlos, 2009).

5.2. *Application à un nouveau corpus : analyse du corpus* EUROPARL

Pour valider l'approche d'acquisition, nous avons cherché à analyser un nouveau corpus afin de déterminer dans quelle mesure la méthode décrite permet d'acquérir de nouveaux SSC en fonction du corpus considéré.

Pour ce faire, nous avons choisi de prendre comme source la partie française du corpus EUROPARL[20] (Koehn, 2005). Il s'agit d'un corpus parallèle librement disponible sur Internet, constitué des actes du Parlement européen entre mars 1996 et septembre 2003[21]. Même s'il ne s'agit pas d'un corpus spécialisé au sens propre du terme, on peut s'attendre au sein du corpus EUROPARL à avoir affaire à des SSC particuliers liés à la nature et au genre du corpus considéré.

Nous comparons dans ce qui suit les résultats obtenus sur le corpus LM10 avec ceux obtenus sur le corpus EUROPARL. Le repérage des nouveaux SSC est simple, dans la mesure où il suffit de les extraire du lexique par comparaison avec LEXSCHEM. On obtient les résultats suivants : pour les 1 934 verbes communs aux deux lexiques, 3 448 couples verbe-SSC sont présents dans les deux ressources, 893 couples sont présents uniquement dans LEXSCHEM-LM10 et 595 uniquement dans LEXSCHEM-EUROPARL, ce qui est déjà important en soi. Ces données sont obtenues entièrement automatiquement.

19. LG a par ailleurs une vaste couverture des constructions pronominales intrinsèques (*s'évanouir, s'imposer*).
20. `http://www.statmt.org/europarl/`.
21. Le corpus inclut onze langues européennes : français, italien, espagnol, portugais, anglais, néerlandais, allemand, danois, suédois, grec et finnois. Chaque langue comprend environ 1 million de phrases, qui contiennent de l'ordre de 28 millions de mots ; nous n'utilisons ici que la partie française du corpus.

Une étude qualitative a ensuite été menée. Parmi les SSC nouveaux repérés, on constate un grande nombre de constructions qui, bien qu'elles ressortissent à la langue générale, n'en sont pas moins remarquables dans le corpus EUROPARL. Ainsi, *enchaîner* se construit fréquemment avec les prépositions *avec* ou *sur*. Ceci est bien évidemment dû au contexte du Parlement européen fait de longues séries de débats et de discussions (*J'enchaînerai immédiatement sur le thème évoqué par M. Ilgenfritz , celui des régions frontalières ; J'enchaînerai ensuite avec le rapport Purvis*). Dans le corpus LM10, *enchaîner* est le plus souvent utilisé au style direct (*« Il n' y a rien de plus excitant que de faire une pièce », enchaîne Martin Gousset.* ou *Elles enchaînent : « Non à l'amnistie »*), alors que cet emploi est quasi inexistant dans le corpus EUROPARL. Dans les deux corpus, le sens concret de *(s')enchaîner* est lui aussi quasi absent.

On constate également l'apparition de nouveaux SSC et de familles sémantiques associées au contexte particulier du corpus EUROPARL. Ainsi, les constructions de *voter* avec les prépositions *pour, contre* ou la locution *en faveur de* sont sans commune mesure avec ce que l'on obtient à partir de LM10 (où seul *pour* est assez présent pour être conservé malgré le filtrage). On a, dans le même ordre d'idées, le verbe *se prononcer* qui peut se construire avec *sur* mais surtout *en faveur de* ; *mettre à disposition* et *donner mandat* qui se construisent avec *pour* ; *légiférer* et *s'abstenir* avec *sur*, etc. Tous ces verbes sont remarquables du domaine considéré et esquissent, en filigrane, les éléments d'un sous-langage législatif.

5.3. *Typologie des erreurs du système ASSCI*

L'évaluation manuelle décrite ci-dessus a également permis de dégager les causes d'erreurs les plus fréquentes dans LEXSCHEM et d'envisager des solutions pour éviter ces erreurs.

Nous avons choisi d'utiliser le couple TreeTagger-SYNTEX pour l'analyse syntaxique car ces outils étaient les plus performants lorsque cette étude a été lancée, au moins sur les corpus de presse (*cf.* section 3.2). Ces outils sont cependant à l'origine de certaines erreurs d'analyse. Certaines erreurs dues au TreeTagger ont été corrigées par un module de post-traitement. On a ainsi pu diminuer drastiquement la proportion de noms propres étiquetés comme verbe (ainsi *Luis* n'est plus reconnu comme une forme du verbe *luire*).

Pour l'analyse syntaxique, SYNTEX a une stratégie prudente : si l'analyseur ne trouve pas d'indice suffisamment fiable pour le rattachement d'un complément, celui-ci peut être laissé « libre », c'est-à-dire qu'il flotte et qu'il n'est rattaché à aucun élément de la phrase (Bourigault *et al.*, 2005). C'est notamment le cas dans certaines phrases comprenant des insertions, comme par exemple *« Il commande ensuite, sur Internet, des pièces détachées, qui donneront une arme parfaitement inutilisable. »* qui produit le schéma (erroné) `[SUJ:SN]` (soit verbe intransitif). Il arrive également que des pronoms ne soient pas rattachés au verbe par l'analyseur. Par exemple, la

phrase *Tu couches ou je te vire* produit le schéma [SUJ:SN] (intransitif) pour le verbe *virer*.

Nous avons partiellement répondu à ce problème en mettant un seuil plus élevé pour le SSC [SUJ:SN], fréquemment produit à cause de ces erreurs d'analyse. Bien évidemment, cette stratégie n'est pas toujours suffisante.

Un dernier ensemble d'erreurs est lié au système d'acquisition lui même. Certains SSC sont incorrects parce qu'ils contiennent des modifieurs. Par exemple, le schéma [SUJ:SN_P-OBJ:SP<dans+SN>] est très présent en corpus pour le verbe *dormir* mais le complément introduit par la préposition *dans* correspond toujours à un complément circonstanciel de lieu : « *Il dort dans son lit.* ». La distinction entre argument et modifieur reste donc difficile quand on se fonde uniquement sur des indices de surface. Notons toutefois que ce type d'erreurs (et plus généralement la présence régulière de certains types de modifieurs) est utile pour le calcul de classes syntaxico-sémantique de verbes.

6. Production de classes de verbes sur la base de leur comportement syntaxique

Les travaux de Levin sur l'anglais (Levin, 1993) tout comme ceux de Gross sur le français (Gross, 1975), malgré leurs différences, tendent à montrer que des verbes partageant des comportements syntaxiques similaires peuvent (souvent) former des classes homogènes sur le plan sémantique. Même si cela ne se vérifie pas dans tous les cas[22], l'intérêt de ces classes est manifeste pour l'élaboration d'une ressource structurée « à la Verbnet », où les verbes sont assemblés en classes syntaxico-sémantiques rangées hiérarchiquement. Il a été montré qu'une approche automatique du type de celle que nous avons présentée constitue une base intéressante pour produire des classes pertinentes et peut notamment enrichir un travail manuel (Kipper *et al.*, 2008).

6.1. *Travaux antérieurs*

Dans sa classification des verbes anglais, Levin (1993) a essayé d'établir un lien entre les alternances syntaxiques et les composants sémantiques qui caractérisent ces classes en anglais. L'intérêt de la classification automatique est d'une part de faciliter la tâche fastidieuse de création de ressources linguistiques par le biais d'une acquisition automatique de propriétés lexicales ; d'autre part, les ressources lexicales structurées sont plus faciles à maintenir et à élargir parce qu'elles permettent de formuler des généralisations sur des classes de mots.

22. Il nous semble d'ailleurs que le statut de ces classes n'est pas tout à fait clair sur le plan théorique ; nous avons, parallèlement au travail présenté ici, entamé un travail de comparaison entre les approches de Levin et de Gross, dans la mesure où il s'agit de deux ensembles de travaux proches et importants (même s'ils reposent sur des présupposés en partie différents) mais qui n'ont jamais fait l'objet d'un examen comparé approfondi. Sur le français, voir aussi les travaux de P. Saint-Dizier, dans la lignée de Levin (Saint-Dizier, 2003).

Les premières tentatives de classification sémantique automatique s'appuyaient sur la classification de Levin : elles visaient à reproduire automatiquement les classes anglaises ou une classification équivalente pour d'autres langues (Schulte im Walde, 2000 ; Korhonen *et al.*, 2003), à identifier les alternances syntaxiques (McCarthy, 2001) ou à compléter le système de Levin par de nouvelles classes (Korhonen et Briscoe, 2004).

Nous nous intéressons ici uniquement aux méthodes non supervisées, reposant sur un espace de traits extrait d'un corpus analysé syntaxiquement (Schulte im Walde, 2000 ; Schulte im Walde et Brew, 2002 ; Korhonen *et al.*, 2003). Le point de départ nécessaire est donc un lexique syntaxique avec des informations concernant la fréquence relative des différents SSC par verbe, avec ou sans information sémantique. Malgré la grandeur des corpus considérés et les informations sur les différents SSC, ces expérimentations montrent que des distinctions syntaxiques plus détaillées ainsi que la prise en compte des modifieurs augmentent la précision de la classification. Cependant, Schulte im Walde (2000) conclut que l'ajout d'informations sur les restrictions de sélection conduit à une problème de manque de données (*data sparseness*) et à la baisse de performance : pour Schulte im Walde, de meilleurs résultats sont obtenus sur des espaces de traits limités à la spécification syntaxique. À l'inverse, (Alishahi et Stevenson, 2007) et (Li et Brew, 2008) ont essayé d'enrichir l'espace de traits de manière efficace par le biais d'informations sémantiques ou lexicales.

Une première tentative pour établir une classification les verbes français est décrite dans (Falk, 2008), qui se fonde sur trois lexiques de sous-catégorisation pour comparer le comportement syntaxique des verbes. La particularité de l'approche est de partir uniquement de ressources manuelles, et donc de ne pas prendre en considération la fréquence des SSC : le calcul de similitude entre les verbes est fait par l'analyse formelle de concepts (Ducassé et Ferré, 2009). Il semble pourtant que les indications concernant la fréquence relative des SSC et la présence (ou non) de modifieurs soient des paramètres importants, aussi prenons-nous en compte ces éléments dans la méthode qui suit.

6.2. *Méthode*

Dans le cadre de cette expérimentation, nous nous sommes intéressés à la classification automatique de verbes français en classes lexico-sémantiques. Pour ce faire, nous partons de l'hypothèse qu'il est possible de produire ce type d'information par des méthodes statistiques appliquées à des données syntaxiques. L'hypothèse est fondée sur l'observation qu'il existe une corrélation entre les propriétés sémantiques des verbes et leurs contextes syntaxiques. Levin (1993) a donné une description systématisée du phénomène en établissant un lien entre les alternances syntaxiques caractéristiques de certains groupes de verbes et les composants sémantiques qui en sont responsables.

Dans la présente étude, nous cherchons à démontrer 1) que la classification automatique de verbes français selon leur comportement distributionnel aboutit souvent à des classes de verbes sémantiquement liés, 2) que les SSC extraits de corpus constituent une représentation fiable de la distribution syntaxique des verbes. Notre objectif est de mettre en place un algorithme de classification aussi général que possible, applicable à de nouveaux verbes, sous condition d'avoir une quantité suffisante d'occurrences dans le corpus pour créer un modèle fiable de leur distribution.

Notre approche est non supervisée, c'est-à-dire que nous ne fournissons au système aucune connaissance autre que les couples verbes-SSC non filtrés issus de l'analyse précédente (section 4.2) avec les informations de fréquence relative qui ont ici une importance cruciale. Nous utilisons en entrée une version non filtrée de la ressource pour deux raisons : le filtrage peut être lui-même une source d'erreurs et, plus fondamentalement, la présence régulière de modifieurs particuliers avec certains verbes est un critère très pertinent pour le calcul des classes sémantiques.

Un léger filtrage est effectué lors de la première étape de la classification, dans le but de réduire l'espace de traits : les schémas qui ont moins de 5 occurrences parmi les verbes à classifier sont exclus. Le nombre de SSC différents – la taille de l'espace de traits – dépend ainsi du vocabulaire verbal utilisé dans l'expérimentation. Dans le cadre de cette expérimentation, nous avons travaillé avec un espace de traits composé de 433 SSC[23]. La représentation des verbes correspond à leur distribution sur tous les schémas considérés dans l'expérimentation (calculée par l'estimation du maximum de vraisemblance, à partir des données de LEXSCHEM) :

$$p(t|v) = f(v,t) / f(v)$$

où f(v) correspond à la fréquence du verbe, et f(v,t) à la fréquence du verbe avec le schéma.

Nous avons utilisé une méthode de regroupement (*clustering*) ascendante hiérarchique. Au début du processus, chaque verbe constitue un groupe à un seul élément (*cluster* – dans ce qui suit, on distingue la notion de « groupe », c'est-à-dire un regroupement obtenu automatiquement, de celle de « classe » correspondant à la référence élaborée manuellement). Lors de chaque itération, les deux groupes de verbes les plus similaires sont unifiés. Cette méthode produit un partitionnement, c'est-à-dire des groupes disjoints de manière à ce que chaque élément à classifier n'appartienne qu'à un seul groupe (*hard clustering*). Bien que cette approche ne permette pas de traiter la polysémie, nous l'avons tout de même choisie pour la facilité de l'interprétation qu'elle offre[24].

23. Puisque le lexique non filtré constitue l'entrée du processus, le nombre des SSC utilisés dépasse celui des SSC dans LEXSCHEM.

24. Une classification des verbes incorporant les problèmes de polysémie serait évidemment souhaitable, mais ceci reste un problème ouvert pour le traitement des langues. Sur le plan pratique, un travail manuel reste également nécessaire si l'on souhaite obtenir un résultat tout à fait fiable. Voir (Kipper *et al.*, 2008) pour une expérience en ce sens, montrant les avantages d'une approche mixte, automatique puis manuelle.

Les distributions ont été comparées avec trois mesures de similarité différentes :

– la divergence de Kullback-Leibler

$$D_{KL}(x \| y) = \sum_{i=1}^{n} x_i \cdot \log \frac{x_i}{y_i} \quad [1]$$

– la divergence de Jensen-Shannon

$$D_{JS}(x \| y) = \frac{1}{2} D_{KL}(x \| M) + \frac{1}{2} D_{KL}(y \| M) \quad [2]$$

où

$$M = \frac{1}{2}(x + y) \quad [3]$$

– et la divergence oblique (*skew divergence*)

$$D_{\alpha}(x \| y) = D_{KL}(x \| \alpha y + (1 - \alpha) x) \quad [4]$$

L'inconvénient de la divergence de Kullback-Leibler est de prendre une valeur indéfinie lorsque la probabilité *y(i)* est 0. Aussi une méthode de lissage simple a-t-elle été appliquée aux données : si la fréquence de cooccurrence du SSC avec le verbe dans le corpus égale zéro, cette valeur sera remplacée par 0,0001, donnant une estimation approximative de la fréquence relative (0,0001 / f(V), où f(V) est la fréquence observée du verbe). La divergence de Jensen-Shannon ainsi que la divergence oblique sont des variantes fondées sur la divergence de Kullback-Leibler qui évitent le problème des valeurs indéfinies par approximation de la valeur de la divergence de Kullback-Leibler. La divergence de Jensen-Shannon est la seule mesure symétrique – pour les autres mesures, le minimum de la distance a été considéré pour chaque paire de verbes comparés. La divergence oblique est une variante pondérée de la divergence de Kullback-Leibler proposée par (Lee, 2001). La pondération se fait par le paramètre libre *a*, dont la valeur optimale est proche de 1 : nous l'avons fixée à 0,99.

Aucune présupposition concernant le nombre et la cardinalité des classes de verbes n'a été incorporée dans l'algorithme. Ainsi, le point d'arrêt du processus de classification dépend de deux paramètres : la distance maximale entre les centres des deux classes à être unifiées et la cardinalité des classes. Ce double paramétrage permet d'éviter l'effet de chaîne, c'est-à-dire le phénomène d'absorption de beaucoup de verbes par quelques groupes très nombreux. Des expérimentations ont été conduites avec des paramètres différentes, et les valeurs optimales de la distance et de la cardinalité maximales ont été établies individuellement pour chaque mesure de distance lors des exécutions de test.

6.3. *Évaluation*

Schulte im Walde propose deux approches différentes pour évaluer une classification automatique (Schulte im Walde, 2009) :

1) mesurer la cohérence à l'intérieur des groupes de verbes obtenus, par une mesure de similarité indépendante de celle utilisée pour la tâche de classification même ;

2) comparer le résultat à une classification (manuelle) de référence.

Dans le cadre de notre expérimentation, nous visons à confirmer l'hypothèse qu'il existe un lien entre le comportement syntaxique des verbes et leurs propriétés sémantiques. Il ne suffit donc pas de démontrer que notre algorithme arrive à modéliser correctement les similarités distributionnelles entre les verbes : c'est la cohérence sémantique des classes qui doit être examinée par comparaison à la référence.

Pour ce faire, nous avons créé à la main une classification dite « de référence ». La référence est composée de 176 verbes, classés dans 16 classes différentes, qui ont d'abord été définies à partir de la classification de Levin, par traduction des verbes anglais. Pour assurer l'homogénéité des classes françaises et une certaine cohérence par rapport au travail de Levin, nous avons vérifié que tous les verbes d'une même classe partageaient un certain nombre de constructions similaires fondamentales. Les classes de la référence sont donc caractérisées par un composant sémantique ainsi que par (au moins) une structure syntaxique en commun. Nous avons enfin vérifié la validité de ces classes en ayant recours au lexique-grammaire, et nous avons pu constater qu'en général les verbes d'une même classe se situent dans la même table du LG, à quelques exceptions près (ce qui montre le besoin d'une comparaison en profondeur des approches et des ressources – comparaison qui sort du cadre de cet article mais que nous menons en parallèle). Pour pouvoir évaluer l'extensibilité et la robustesse de la méthode, des verbes de fréquences différentes ont été inclus dans l'expérimentation. La cardinalité des classes varie entre 8 et 17.

Les résultats ont été évalués par rapport à la référence selon quatre mesures. La difficulté de la tâche de classification dépend du nombre de classes. Pour une classification à *m* classes, la valeur basse (*baseline*) de l'exactitude (*accuracy*) est de *1/m*, soit 0.0625 dans notre cas.

Mesure de distance	Card.	APP	mPURITY	ACC	F-measure
KL	6	0,13	0,48	0,30	0,36
KL	5	0,13	0,51	0,27	0,35
JS	4	0,21	0,60	0,28	0,39
JS	5	0,18	0,54	0,30	0,38
skew	4	0,22	0,62	0,27	0,37
skew	5	0,18	0,55	0,29	0,37
skew	9	0,16	0,47	0,35	0,40

Les groupes de verbes résultants ont été comparés à la référence d'une part par la mesure *Adjusted Pairwise Precision*, qui calcule la précision des groupes en prenant

les verbes d'une même classe deux à deux (puis en comparant la sortie du système avec la référence pour vérifier s'ils appartiennent bien à la même classe dans les deux cas), et en prenant en compte la cardinalité (afin de pénaliser les petits groupes de verbes) :

$$APP\,(C) = \frac{1}{|C|} \sum_{i=1}^{C} \frac{paires_correctes_dans_c_i}{paires_dans_c_i} \times \frac{c_i - 1}{c_i + 1} \qquad [5]$$

D'autre part, il est possible d'associer les groupes de verbes résultant de la classification non supervisée aux classes de la référence en établissant la correspondance selon la classe sémantique prédominante à l'intérieur du groupe. Cela nous permet de calculer la pureté modifiée (*modified purity* – pureté moyenne des classes) et l'exactitude pondérée de classes (*weighted class accuracy* – rappel pondéré en fonction de la taille des classes dans la référence) (Korhonen *et al.*, 2008). Lors du calcul de la pureté modifiée, les éléments qui n'appartiennent pas à la classe prédominante, ainsi que les singletons sont considérés comme des erreurs.

$$mPurity\,(C) = \frac{\sum_{n_{prevalent}(k_i) \geq 2} n_{prevalent}\,(k_i)}{|C|} \qquad [6]$$

L'exactitude pondérée des classes peut être considérée comme une mesure de rappel : pour chaque classe de la référence, elle considère la quantité des verbes appartenant au groupe dominant associé à cette classe. Par définition, cette quantité ne peut pas dépasser la cardinalité maximale des groupes.

$$Acc\,(C) = \frac{\sum_{i=1}^{C} verbes_dans_GRP.DOM_i}{|C|} \qquad [7]$$

La F-mesure a été calculée avec des poids égaux pour le rappel et la précision :

$$F = \frac{2 \times mPurity \times Acc}{mPurity + Acc} \qquad [8]$$

En optimisant les paramètres pour la mesure APP, les groupes à quatre éléments donnent les meilleurs résultats. La précision forte semble soutenir le lien supposé entre les propriétés sémantiques et la distribution syntaxique observée dans le corpus, comme dans les exemples suivants :

```
groupe : errer voyager circuler naviguer
groupe : dire indiquer affirmer déclarer
groupe : signaler révéler montrer annoncer
groupe : ressentir définir désigner percevoir
groupe : rouspéter ronchonner grogner râler
```

Il est important de noter que cette qualité de la classification (20 % des classes sont parfaitement homogènes, 43 % contiennent 1 verbe incorrect au maximum) a été obtenue en utilisant une chaîne de traitement entièrement automatisée, de l'analyse de corpus jusqu'à la construction de l'espace de traits pour la classification. De plus, l'espace de traits est conçu pour être aussi général que possible, n'incorporant aucune connaissance préalable sur la classification de référence.

Cependant, les mesures de rappel pénalisent plus sérieusement la différence structurelle entre la classification résultante et la classification de référence, notamment en ce qui concerne le nombre et la cardinalité des classes. La mesure d'exactitude montre que la cohérence des classes baisse avec l'augmentation de la cardinalité au dessus de 4. En observant les résultats, nous pouvons noter que c'est souvent l'effet de chaîne qui affaiblit la cohérence des groupes de verbes : au lieu de réunir des verbes autour d'un composant sémantique central, ils sont composés d'une série de paires avec un lien sémantique qui se modifie constamment par l'ajout d'un nouvel élément. Par exemple, les groupes ci-dessous affichent une certaine cohérence sémantique, mais la relation sémantique est modifiée par rapport à la référence (les crochets indiquent la classification de référence) :

```
groupe : [resplendir pêtiller scintiller] [vibrer]
groupe : [consterner ennuyer] [dévisager] [rosser]
groupe : [bougonner gémir] [trembler vaciller]
groupe : [grésiller geindre] [trembloter] [flamboyer]
groupe : [consolider renforcer] [réintégrer] [maintenir]
```

Plusieurs améliorations peuvent être envisagées. Une classification supervisée permettrait de spécifier les composants sémantiques centraux des classes, et ainsi d'adapter l'espace de traits (par filtrage ou pondération) à la tâche spécifique. Les schémas de sous-catégorisation les plus spécifiques aux classes donneront une idée des alternances qui caractérisent les classes sémantiques de verbes français. Le recours à d'autres traits (notamment des informations quant aux restrictions de sélection) permettrait d'obtenir des classes différentes et sans doute plus précises que celles obtenues en l'état.

7. Conclusion

Nous avons présenté dans cet article un système d'acquisition de lexique syntaxique pour le français et un système de classification syntaxico-sémantique des verbes reposant sur ce système d'acquisition. Les expériences sur le verbe montrent l'intérêt de notre méthode : le système est notamment capable de repérer, à moindre coût, des données nouvelles afin d'enrichir les lexiques existants. Au-delà, le système permet d'acquérir des données profilées en fonction d'un corpus donné, par exemple pour fournir à un analyseur syntaxique probabiliste des schémas de sous-catégorisation pondérés. Le système d'acquisition peut aussi permettre l'étude contrastive de corpus variés et des expériences sont en cours dans cette direction (afin de voir quelles constructions sont utilisées de façon remarquable dans un corpus donné

par rapport à un autre corpus, etc.). Les travaux sur l'acquisition de classes lexico-sémantiques posent enfin des questions théoriques sur la nature des classes obtenues. Si l'intérêt applicatif de classes sémantiques ne fait pas de doute, il nous semble nécessaire de continuer à s'interroger sur la nature même de ces travaux et sur ce qu'ils nous disent sur la langue.

Remerciements

Nous tenons à remercier les trois relecteurs anonymes de la revue TAL pour leurs remarques pertinentes qui nous ont permis de grandement améliorer la qualité de l'article. Nous remercions également Béatrice Pelletier pour sa relecture attentive.

La thèse de Cédric Messiant a été financée par une allocation DGA. Ces recherches s'inscrivent par ailleurs dans le cadre des projets PHC TAACL (Technologies multilingues pour l'Acquisition Automatique de Connaissances Lexicales) et ANR CroTAL (Conditional RandOm Fields pour le Traitement Automatique des Langues).

8. Bibliographie

Abeillé A., Clément L., Toussenel F., « Building a Treebank for French », *in* A. Abeillé (ed.), *Treebanks : Building and Using Parsed Corpora*, Kluwer Academic Publishers, Dordrecht, p. 165-187, 2003.

Abeillé A., *Les nouvelles syntaxes*, Armand Colin, Paris, 1993.

Agirre E., Edmonds P. (eds), *Word Sense Disambiguation : Algorithms and Applications*, Springer, Berlin, 2007.

Alishahi A., Stevenson S., « A Cognitive Model for the Representation and Acquisition of Verb Selectional Preferences », *ACL Workshop on Cognitive Aspects of Computational Language Acquisition*, Prague, Czech Republic, p. 41-48, 2007.

Borillo A., « Remarques sur les verbes symétriques du français », *Langue française*, vol. 11, p. 17-31, 1971.

Bourigault D., *Un analyseur syntaxique opérationnel : SYNTEX*, Mémoire d'Habilitation, Université de Toulouse-le-Mirail, 2007.

Bourigault D., Jacques M.-P., Fabre C., Frérot C., Ozdowska S., « Syntex, analyseur syntaxique de corpus », *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN)*, Dourdan, 2005.

Brent M. R., « Automatic Acquisition of Subcategorization Frames from Untagged Text », *Proceedings of the Meeting of the Association for Computational Linguistics (ACL)*, Berkeley, CA, p. 209-214, 1991.

Brent M. R., « From Grammar to Lexicon : Unsupervised Learning of Lexical Syntax », *Computational Linguistics*, vol. 19, p. 203-222, 1993.

Bresnan J., Zaenen A., « Deep unaccusativity in LFG », *in* K. Dziwirek (ed.), *Grammatical Relations. A Cross-Theoretical Perspective*, Center for the Study of Language and Information, Stanford University, 1990.

Briscoe T., Carroll J., « Automatic Extraction of Subcategorization from Corpora », *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, Washington, DC., p. 356-363, 1997.

Chesley P., Salmon-Alt S., « Automatic extraction of subcategorization frames for French », *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Gênes, 2006.

Constant M., Tolone E., « A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables », *Actes du 27ème Colloque international sur le Lexique et la Grammaire (LGC'08)*, L'Aquila, Italie, p. 11-18, 2008.

Copestake A., The Representation of Lexical Semantic Information, PhD thesis, University of Sussex, 1992.

Cruse A. D., *Lexical semantics*, Cambridge University Press, Cambridge, 1986.

Danlos L., « Les lexiques en traitement automatique du langage naturel », *Proceedings of the 3rd meeting on Langage Industry*, Grossetto, 1988.

Dendien J., Pierrel J.-M., « Le Trésor de la Langue Française Informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence », *Traitement Automatique des Langues*, vol. 2, p. 11-37, 2003.

Dubois J., Dubois-Charlier F., *Les Verbes français*, Larousse-Bordas, Paris, 1997.

Ducassé M., Ferré S., « Aide à la décision multicritère : cohérence et équité grâce à l'analyse de concepts », *Modèles et Apprentissage en Sciences Humaines et Sociales*, 2009.

Falk I., *Création automatique de classes sémantiques verbales pour le français*, Mémoire de Master, LORIA, Nancy, 2008.

Fort K., Guillaume B., « PrepLex : un lexique des prépositions du français pour l'analyse syntaxique », *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN)*, Toulouse France, 2007.

Francopoulo G., « TagParser et Technolangue-Easy », *Actes de l'Atelier technolangue, TALN'05*, Dourdan, 2005.

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C., « Lexical Markup Framework (LMF) », *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Genoa, 2006.

Gardent C., Guillaume B., Perrier G., Falk I., « Extraction d'information de sous-catégorisation à partir des tables du LADL », *Actes de la conférence Traitement Automatique des Langues Naturelles*, Louvain, Belgique, 2006.

Gross M., *Méthodes en syntaxe*, Hermann, Paris, 1975.

Gross M., Danlos L., « Building Electronic Dictionaries for Natural Language Processing », *Programming of Future Generation Computers*, North Holland, Elsevier Science Publishers, Amsterdam, 1988.

Guillet A., Leclère C., *La structure des phrases simples en français – constructions transitives locatives*, Droz, Genève-Paris, 1992.

Jackendoff R., *Semantic Structures*, The MIT Press, Cambridge, Massachusetts, 1990.

Kipper K., Korhonen A., Ryant N., Palmer M., « A Large-Scale Classification of English Verbs », *Journal of Language Resources and Evaluation*, vol. 42, n° 1, p. 21-40, 2008.

Koehn P., « Europarl : A Parallel Corpus for Statistical Machine Translation », *MT Summit*, Phuket Island, Thailand, 2005.

Koenig J.-P., Davis A., « Semantically transparent linking in HPSG », *Proceedings of the HPSG03 Conference*, East Lansing, Michigan, p. 222-235, 2000.

Korhonen A., Subcategorization acquisition, PhD thesis, University of Cambridge, 2002.

Korhonen A., Briscoe T., « Extended Lexical-Semantic Classification of English Verbs », *in* D. Moldovan, R. Girju (eds), *HLT-NAACL 2004 : Workshop on Computational Lexical Semantics*, Association for Computational Linguistics, Boston, Massachusetts, USA, p. 38-45, May 2 - May 7, 2004.

Korhonen A., Gorrell G., McCarthy D., « Statistical filtering and subcategorization frame acquisition », *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong, 2000.

Korhonen A., Krymolowski Y., Collier N., « The Choice of Features for Classification of Verbs in Biomedical Texts », *Proceedings of the 2008 COLING conference*, Manchester, p. 449-456, 2008.

Korhonen A., Krymolowski Y., Marx Z., « Clustering Polysemic Subcategorization Frame Distributions Semantically », *Proceedings of the Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, p. 64-71, 2003.

Kupsc A., « Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré », *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN)*, Toulouse, June, 2007.

Laporte E., « Mots et niveau lexical », *in* J.-M. Pierrel (ed.), *Ingénierie des langues*, Hermès, p. 25-49, 2000.

Lee L., « On the Effectiveness of the Skew Divergence for Statistical Language Analysis », *Proceedings of the Workshop on Artificial Intelligence and Statistics*, Key west, Florida, 2001.

Levin B., *English Verb Classes and Alternations : a preliminary investigation*, University of Chicago Press, Chicago and London, 1993.

Levin B., Rappaport Hovav M., *Argument Realization*, Cambridge University Press, Cambridge, 2005.

Li J., Brew C., « Which Are the Best Features for Automatic Verb Classification », *Proceedings of the Meeting of the Association for Computational Linguistics (ACL–HLT)*, Columbus, Ohio, p. 434-442, 2008.

Manning C. D., « Automatic Acquisition of a Large Subcategorization Dictionary from Corpora », *Proceedings of the Meeting of the Association for Computational Linguistics (ACL)*, p. 235-242, 1993.

McCarthy D., *Lexical Acquisition at the Syntax-Semantics Interface : Diathesis Alternations*, PhD Thesis, University of Sussex, 2001.

Mel'cuk I., Polguère A., « Dérivations sémantiques et collocations dans le DiCo/LAF », *Langue française*, vol. 150, p. 66-83, 2006.

Messiant C., « A Subcategorization Acquisition System for French Verbs », *Proceedings of the ACL-08 : HLT Student Research Workshop*, Association for Computational Linguistics, Columbus, Ohio, p. 55-60, June, 2008.

Moreau E., Tellier I., Balvet A., Laurence G., Rozenknop A., Poibeau T., « Annotation fonctionnelle de corpus arborés avec des Champs Aléatoires Conditionnels », *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN)*, Senlis, 2009.

O'Donovan R., Burke M., Cahill A., van Genabith J., Way A., « Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks », *Computational Linguistics*, vol. 31, n° 3, p. 329-366, 2005.

Poibeau T., Messiant C., « Do we still need gold standard for evaluation ? », *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marrakech, 2008.

Preiss J., Briscoe T., Korhonen A., « A System for Large-Scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora », *Proceedings of the Meeting of the Association for Computational Linguistics (ACL)*, Prague, p. 912-918, 2007.

Pustejovsky J., *The generative lexicon*, The MIT Press, Cambridge, 1995.

Sagot B., « The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French », *Language Resource and Evaluation Conference (LREC)*, La Valette, 2010.

Sagot B., Danlos L., « Constructions pronominales dans Dicovalence et le lexique-grammaire – Intégration dans le Lefff », *Linguisticæ Investigationes*, vol. 32, n° 2, p. 293-304, 2009.

Saint-Dizier P., « Quelques défis et éléments de méthode pour la construction de ressources lexicales sémantiques », *Revue Française de Linguistique Appliquée*, vol. 23, p. 34-47, 2003.

Salkoff M., Valli A., « La constitution d'un lexique de la complémentation verbale du français », *Actes du Colloque international sur le lexique et la grammaire*, Palerme, 2006.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *International Conference on New Methods in Language Processing*, unknown, Manchester, UK, 1994.

Schulte im Walde S., « Clustering Verbs Semantically According to their Alternation Behaviour », *Proceedings of COLING*, Saarbrücken, 2000.

Schulte im Walde S., « The Induction of Verb Frames and Verb Classes from Corpora », *Corpus Linguistics. An International Handbook*, Mouton de Gruyter, Berlin, 2009.

Schulte im Walde S., Brew C., « Inducing German Semantic Verb Classes from Purely Syntactic Subcategorisation Information », *Proceedings of the Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, p. 223-230, 2002.

Stevenson S., Carreras X. (eds), *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, ACL, Boulder, Colorado, 2009.

van den Eynde K., Blanche-Benveniste C., « Syntaxe et mécanismes descriptifs : présentation de l'approche pronominale », *Cahiers de Lexicologie*, vol. 32, p. 3-27, 1978.

van den Eynde K., Mertens P., *Le dictionnaire de valence Dicovalence : manuel d'utilisation*, Manuscript, Leuven, 2006.

van Rullen T., Blache P., Portes C., Rauzy S., Maeyheux J.-F., Guénot M.-L., Balfourier J.-M., Bellengier E., « Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales », *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN)*, Dourdan, 2005.